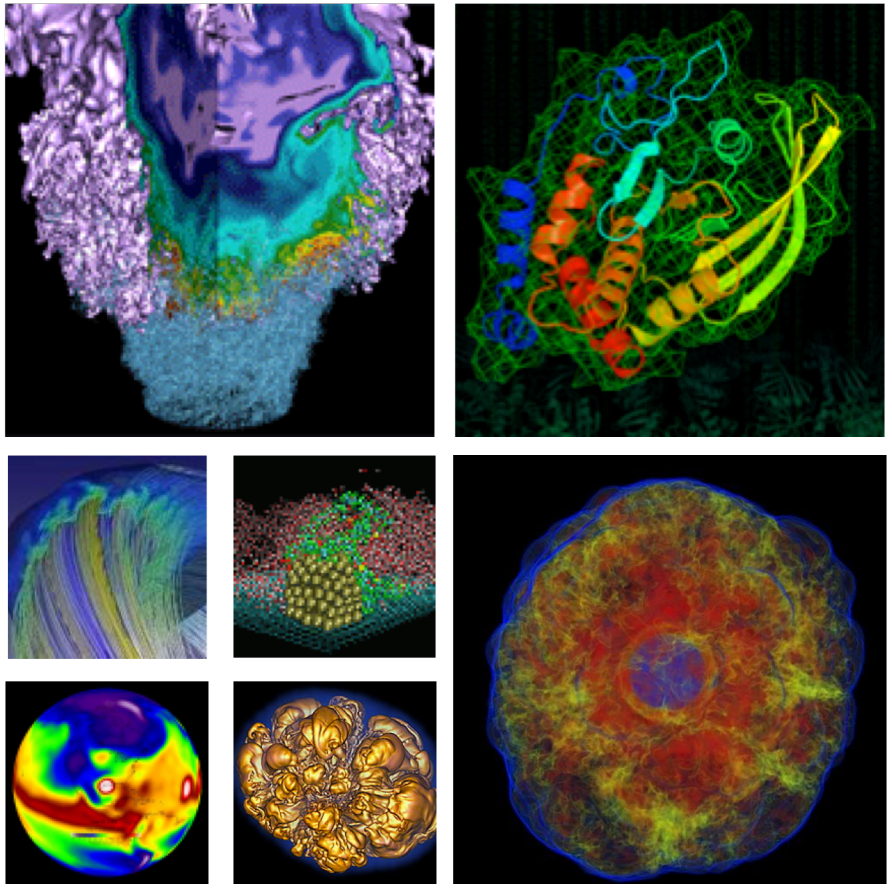
Big Data Center



NERSC

Prabhat
7/18/2018

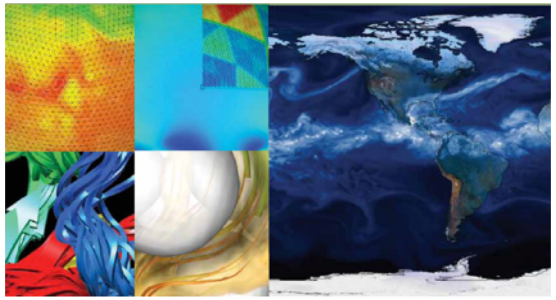
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

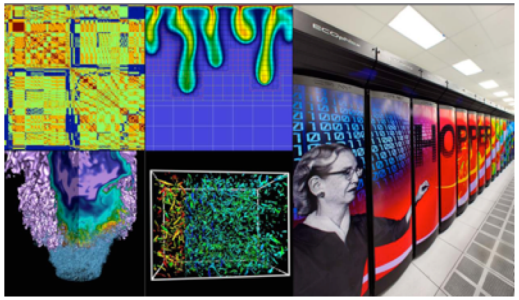
NERSC: the Mission HPC Facility for DOE Office of Science Research

U.S. DEPARTMENT OF ENERGY | Office of Science

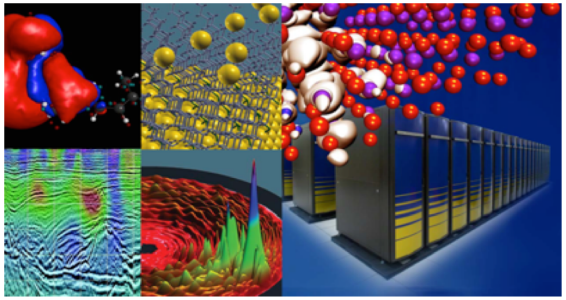
Largest funder of physical science research in the U.S.



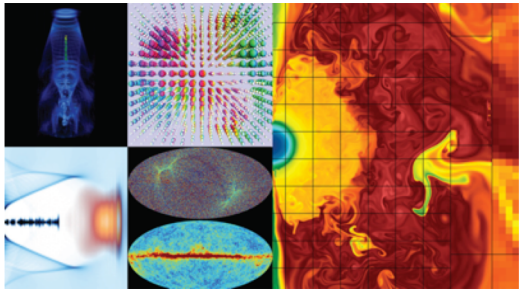
Bio Energy, Environment



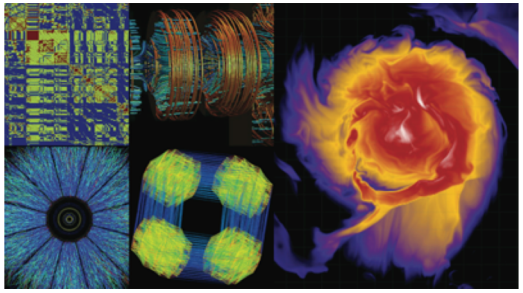
Computing



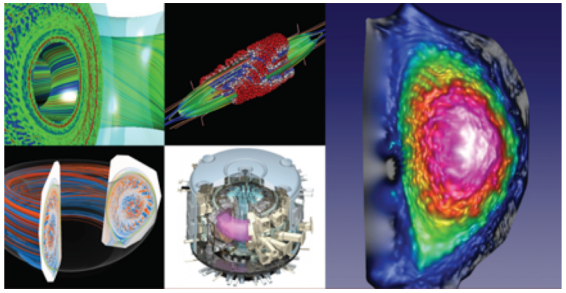
Materials, Chemistry, Geophysics



Particle Physics, Astrophysics



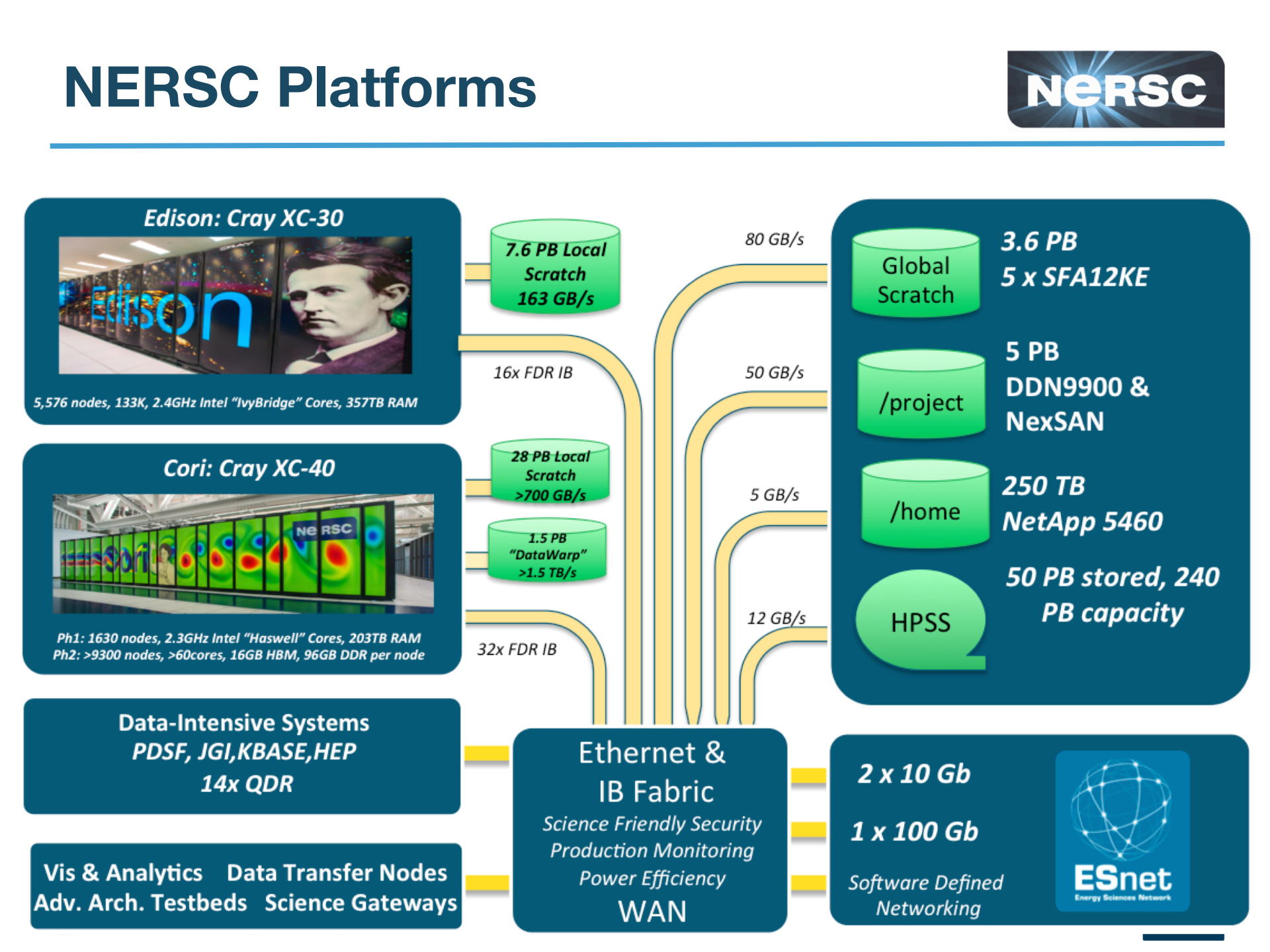
Nuclear Physics



Fusion Energy, Plasma Physics

7,000 users, 700 projects, 700 codes, 48 states, 40 countries, universities & national labs

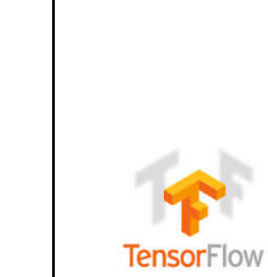
NERSC Platforms



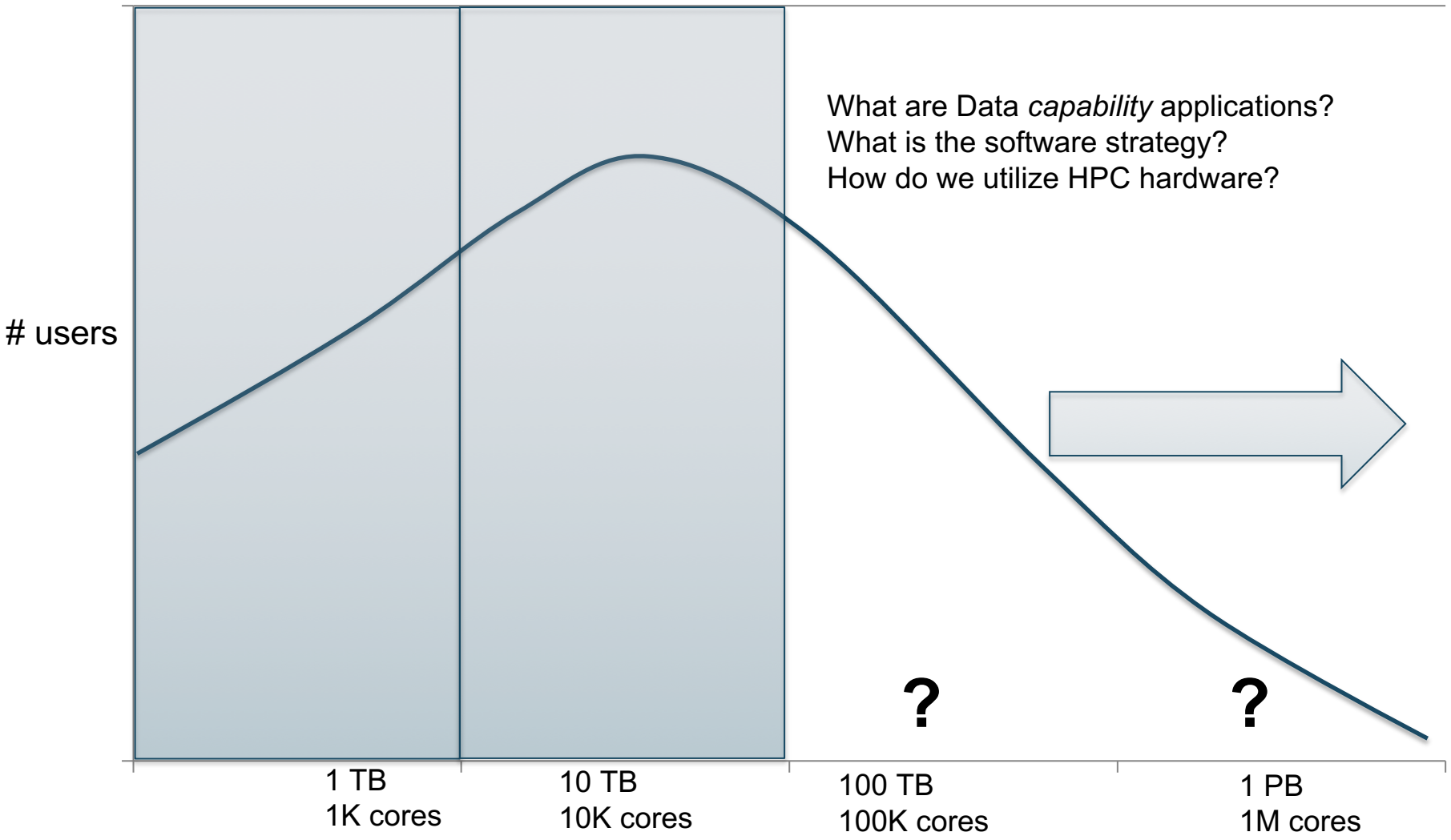
NERSC Big Data Stack

Capabilities	Technologies
Data Transfer + Access	globus online, GridFTP, jupyter, IM, python django, newt
Workflows	FireWorks, taskfarmer
Data Management	HDF, netCDF, ROOT Data Analysis Framework, mongoDB, MySQL, PostgreSQL
Data Analytics	python, Spark, MATLAB, TensorFlow, PYTORCH, R, julia, Caffe, K
Data Visualization	VisIT, ParaView

Deep Learning Stack

	Technologies
Deep Learning Frameworks	 Caffe PYTORCH Theano, Neon, CNTK, MXNet, …
Multi Node libraries	MLSL Cray Plugin Horovod GRPC MPI
Single Node libraries	MKL-DNN CuDNN
Hardware	CPUs (KNL) GPUs FPGAs Accelerators

Big Data Center



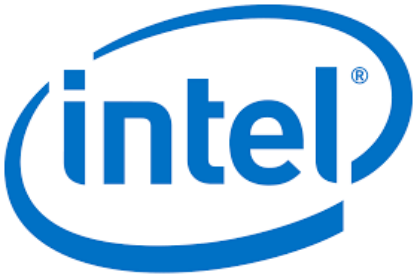
Big Data Center

Mission:

Solve DOE's leading data-intensive science problems at scale on Cori

Performance Optimization and Scaling of Production Data Analytics and Management Technologies

Big Data Center Collaboration



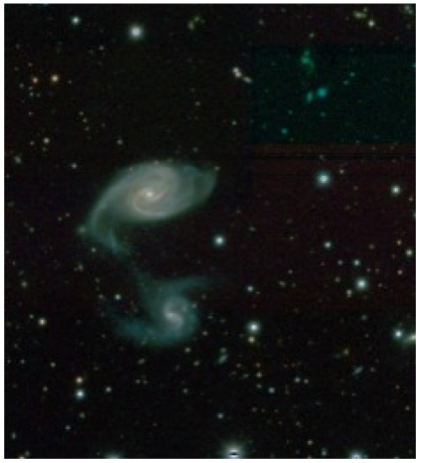
Creating a catalog of all visible objects in the Universe

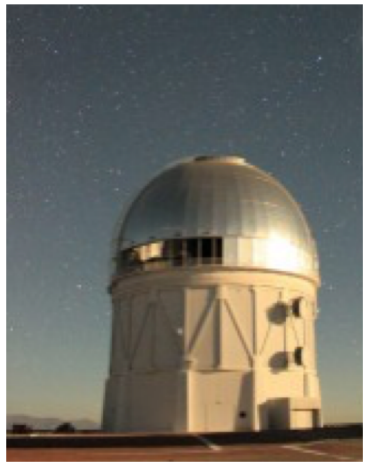
U.S. DEPARTMENT OF ENERGY | Office of Science

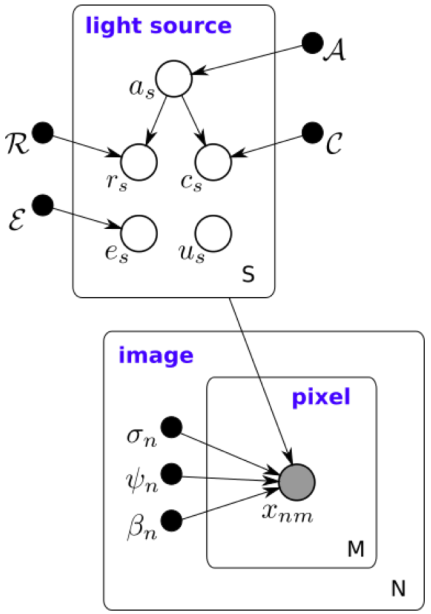
BERKELEY LAB
Lawrence Berkeley National Laboratory

Celeste: 1st Julia application to achieve 1PF

- **Scientific Achievement**
  - First catalog with parameter and *uncertainty* estimates for 188M objects
  - 55 TB SDSS dataset processed in 15 minutes on Cori @ NERSC
  - DESI instrument will use catalog for target selection
- **Statistics Achievement**
  - Bayesian Inference on world's largest generative model (in science)
  - Joint estimation of 8B parameters
- **CS Achievement**
  - Code written in Julia, optimized for execution on KNL
  - Code scaled on 9300 KNL nodes
  - Uses BB for staging data

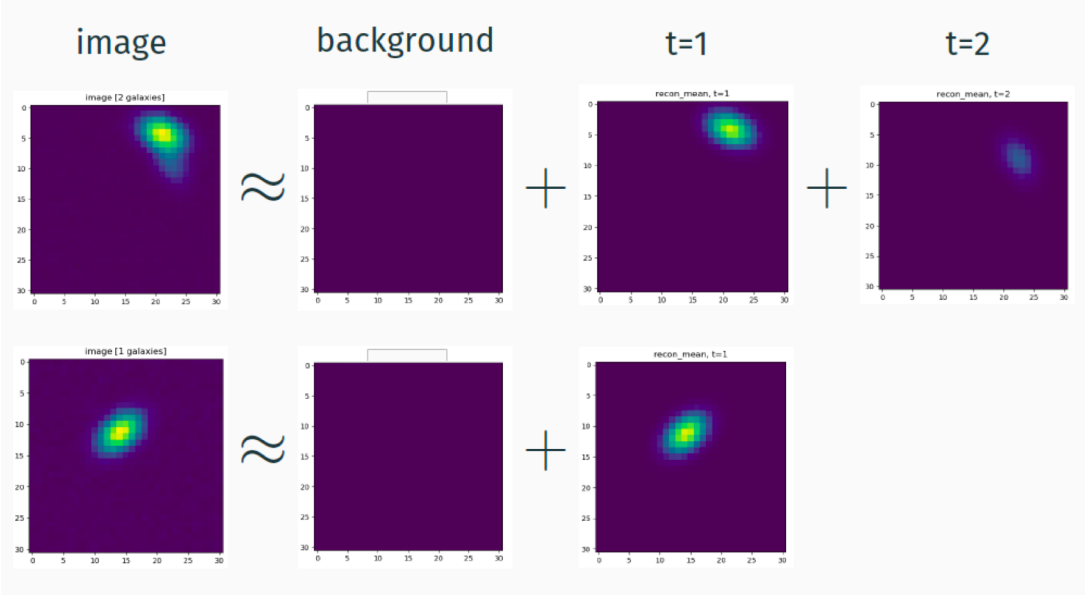






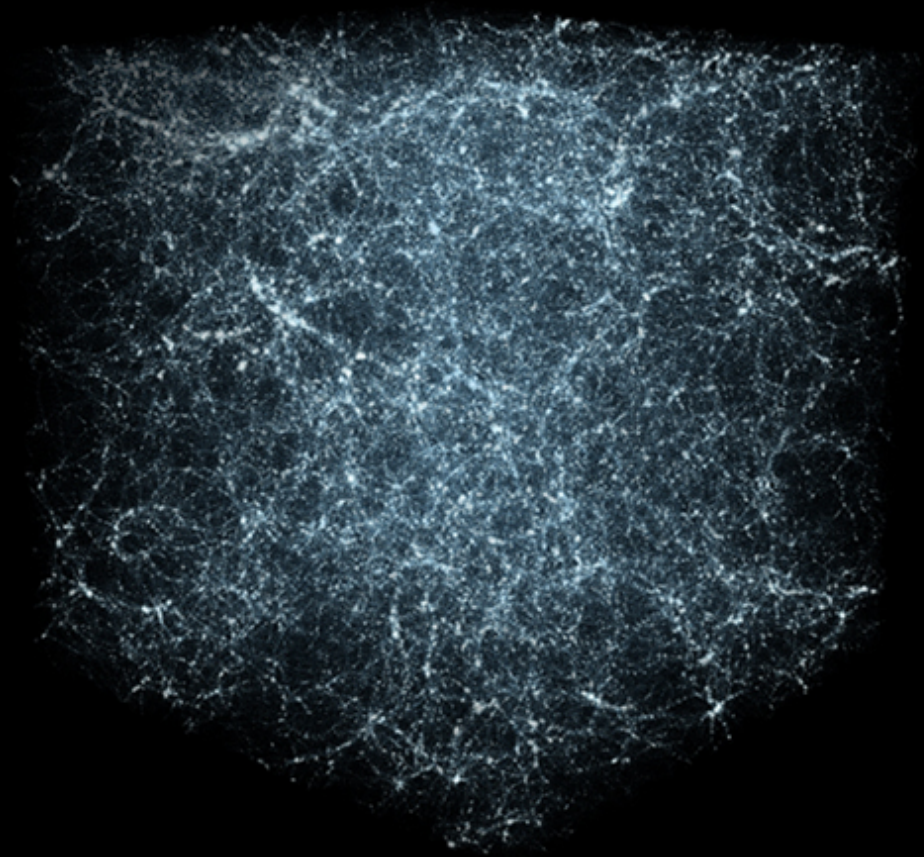
Celeste 2.0: Graphical Model + Deep Learning

- Jeff Regier, Bryan Liu, Jon McAuliffe (UCB); Andy Miller, Ryan Adams (Harvard); David Schlegel (LBL Physics)



- Auto-encoder and RNN for modeling variable number of light sources
- Gravitational Lensing

Determining the Fundamental Constants of Cosmology

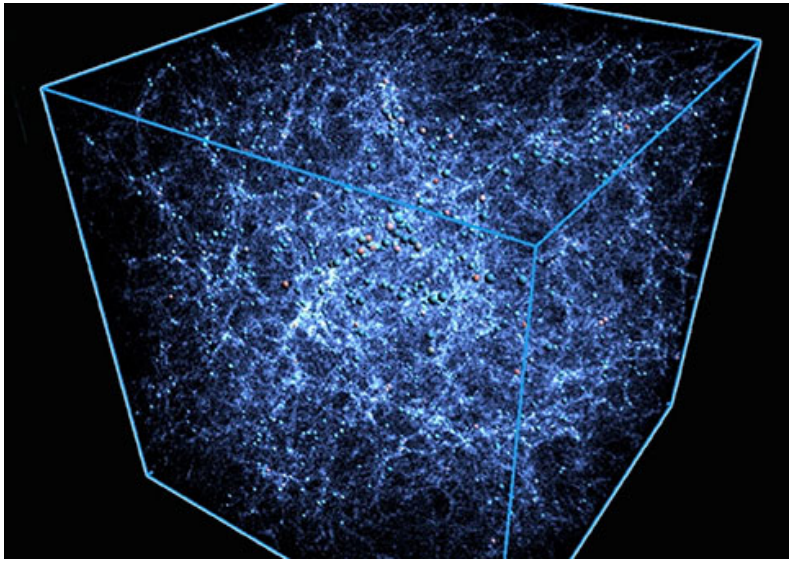


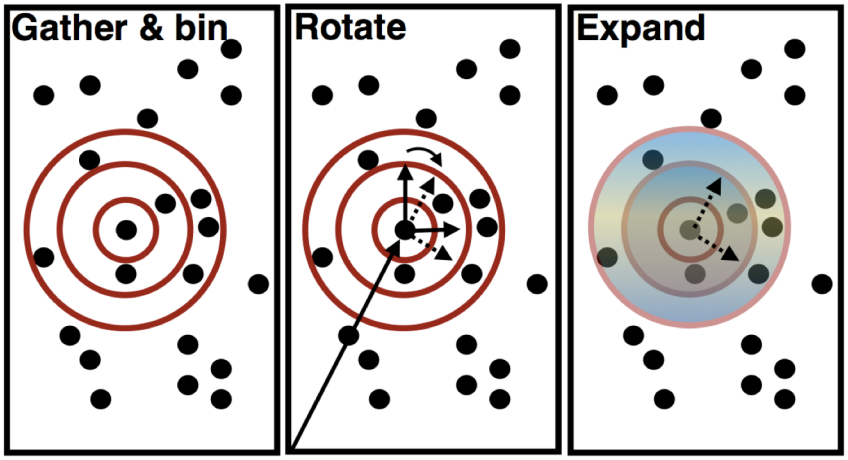
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Galactos

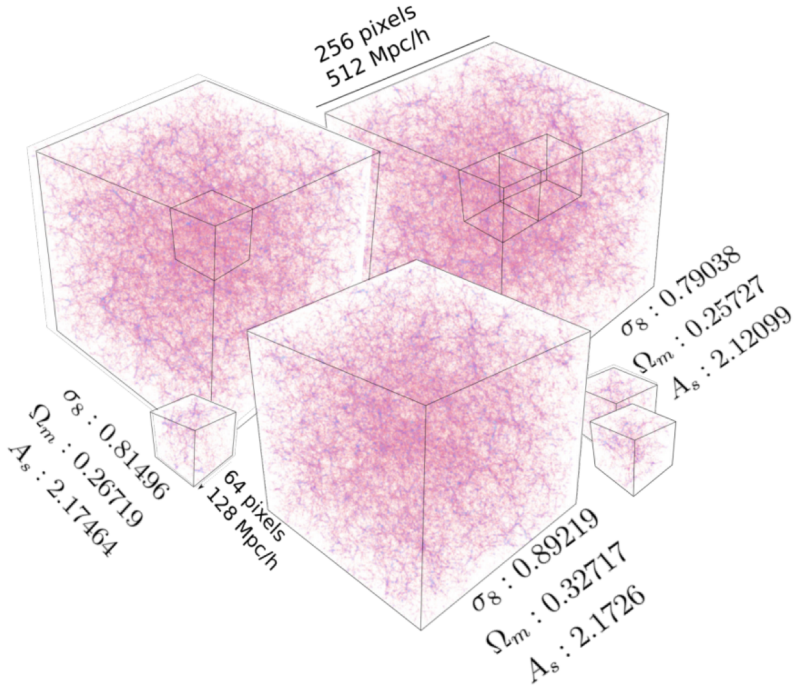
- **Scientific Achievement**
  - Computed 3-pt correlation for the largest scientific dataset (2B galaxies from Outer Rim) in ~20 minutes
  - Code is ready for LSST: 10B galaxies
- **Methods Achievement**
  - Novel O(N^2) algorithm based on spherical harmonics
- **CS Achievement**
  - Optimized KD-tree for spatial partitioning and querying
  - Single node implementation achieves 39% peak (max: 50%)
  - Code achieves 9.8PF (mixed precision)

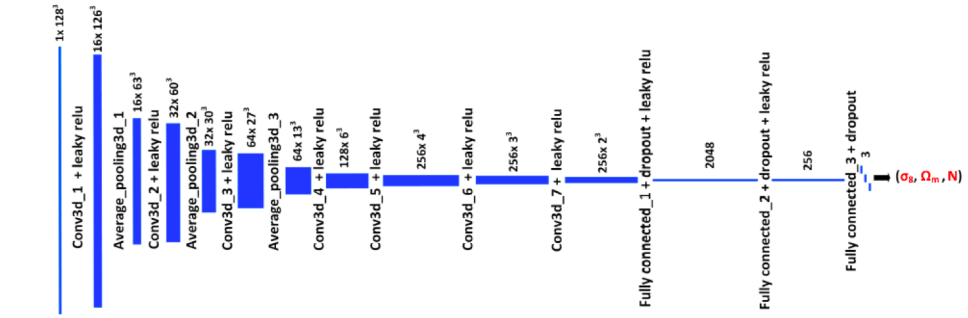


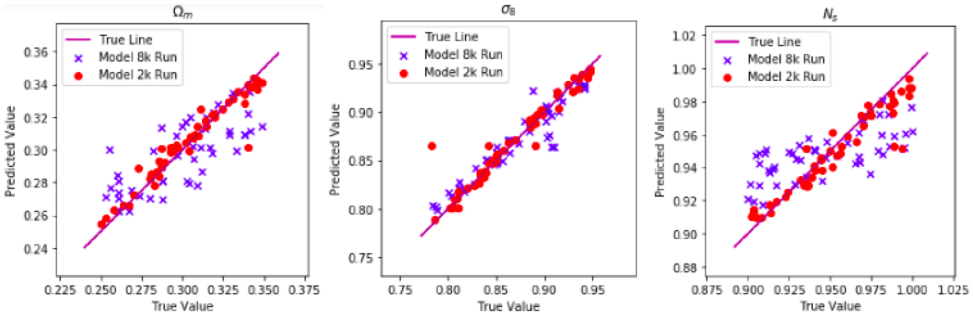


CosmoFlow

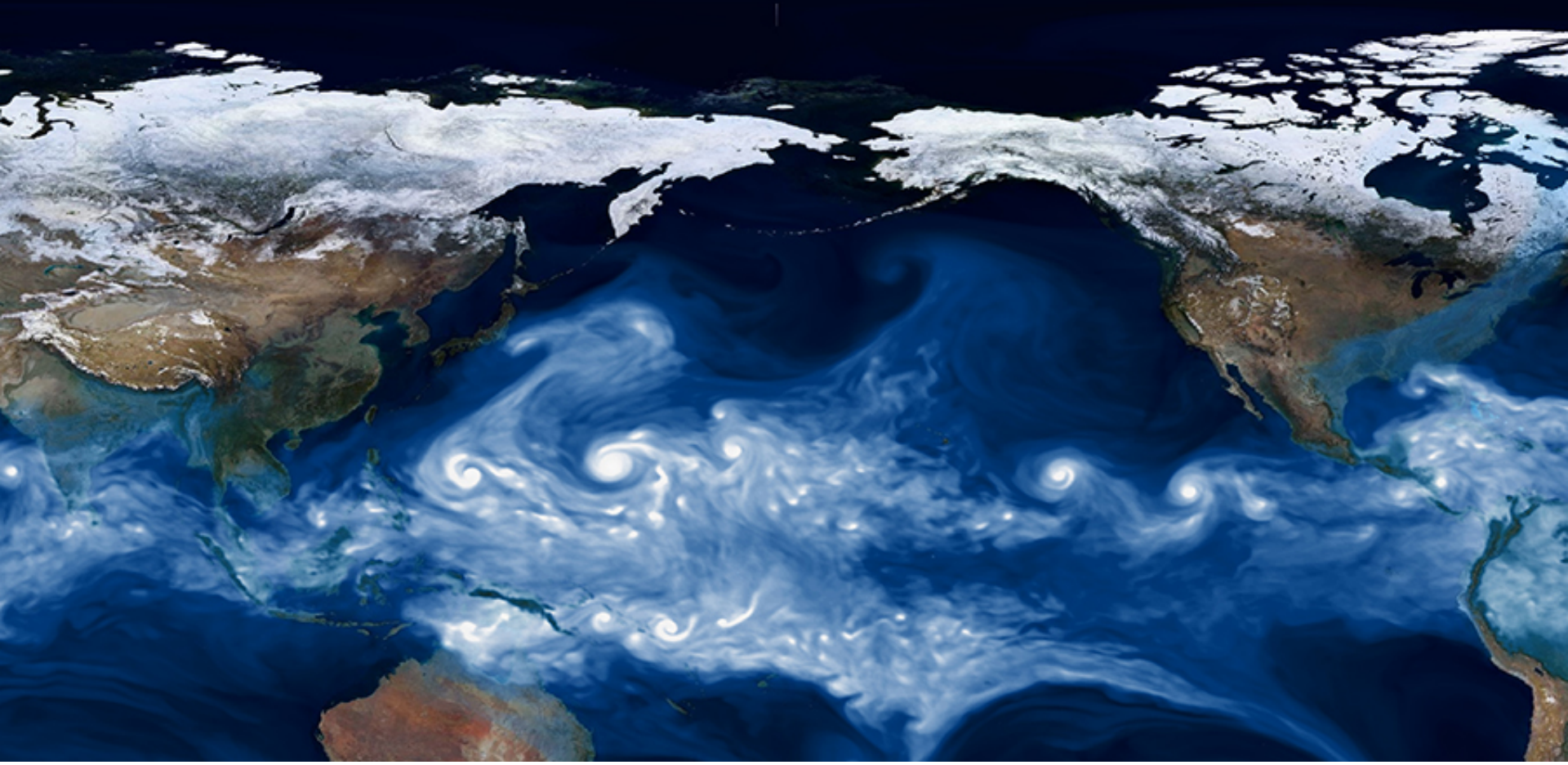
- Amrita Mathuriya (Intel); Debbie Bard (NERSC); Pete Mendygral (Cray); Siyu He, Shirley Ho (LBL Physics); et al.
- 3D CNN implemented in TensorFlow scaled to 8192 Cori/KNL nodes with Cray DL plugin. 3.5 PF peak performance.







Characterizing Extreme Weather in a Changing Climate

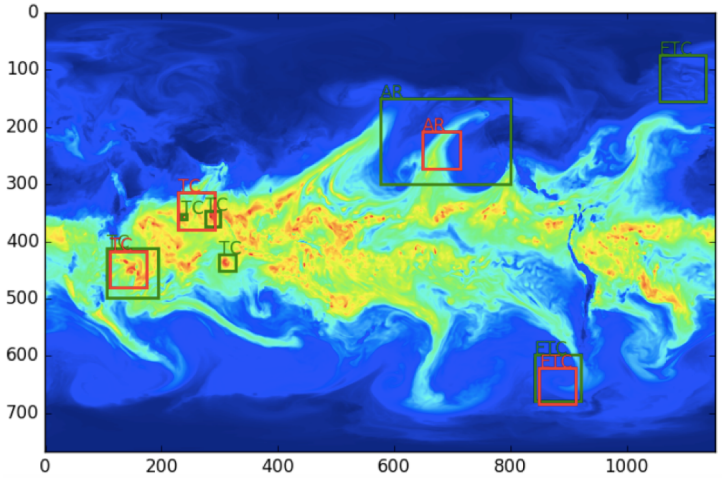


U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

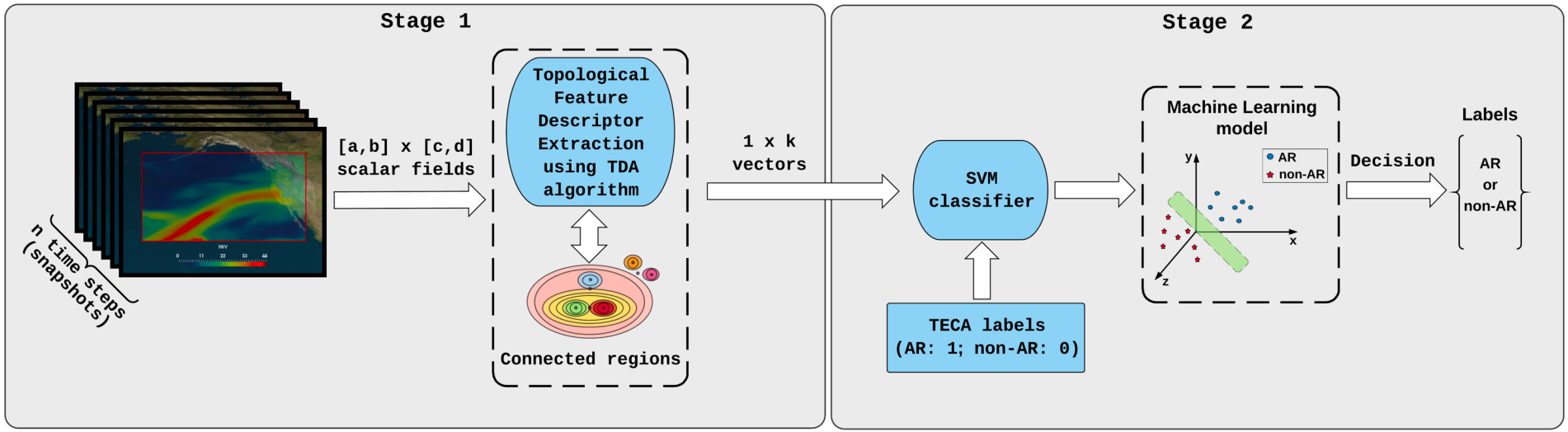
SC'17:: Deep Learning at 15PF

- **Scientific Achievement**
  - Unified architecture for climate pattern classification
- **Methods Achievement**
  - Semi-supervised convolutional architecture
  - Hybrid parameter update strategy
- **CS Achievement**
  - IntelCaffe + MLSL optimized on KNL
  - 2TF peak on single KNL node
  - 15 PF peak on 9300 Cori nodes



Topological Analysis

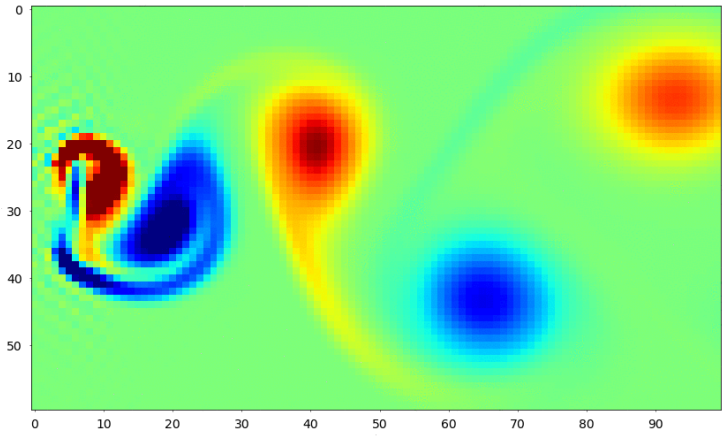
- Vitaliy Kurlin, Grzegorz Muszynski (U. Liverpool); Karthik Kashinath (NERSC)

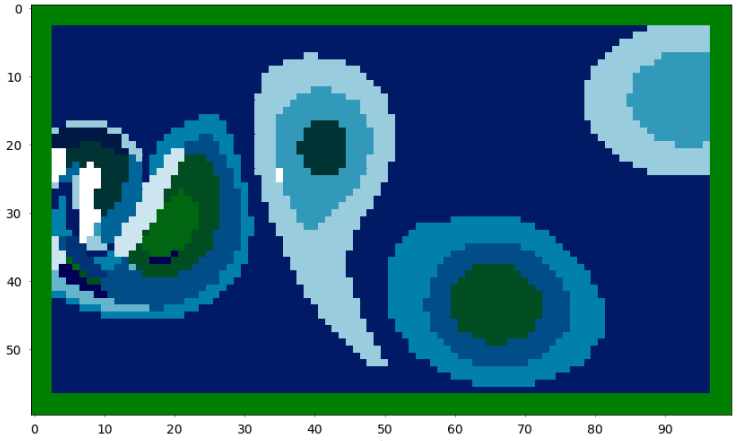


- Method is independent of threshold specification
- Classification accuracy for Atmospheric Rivers: 91%

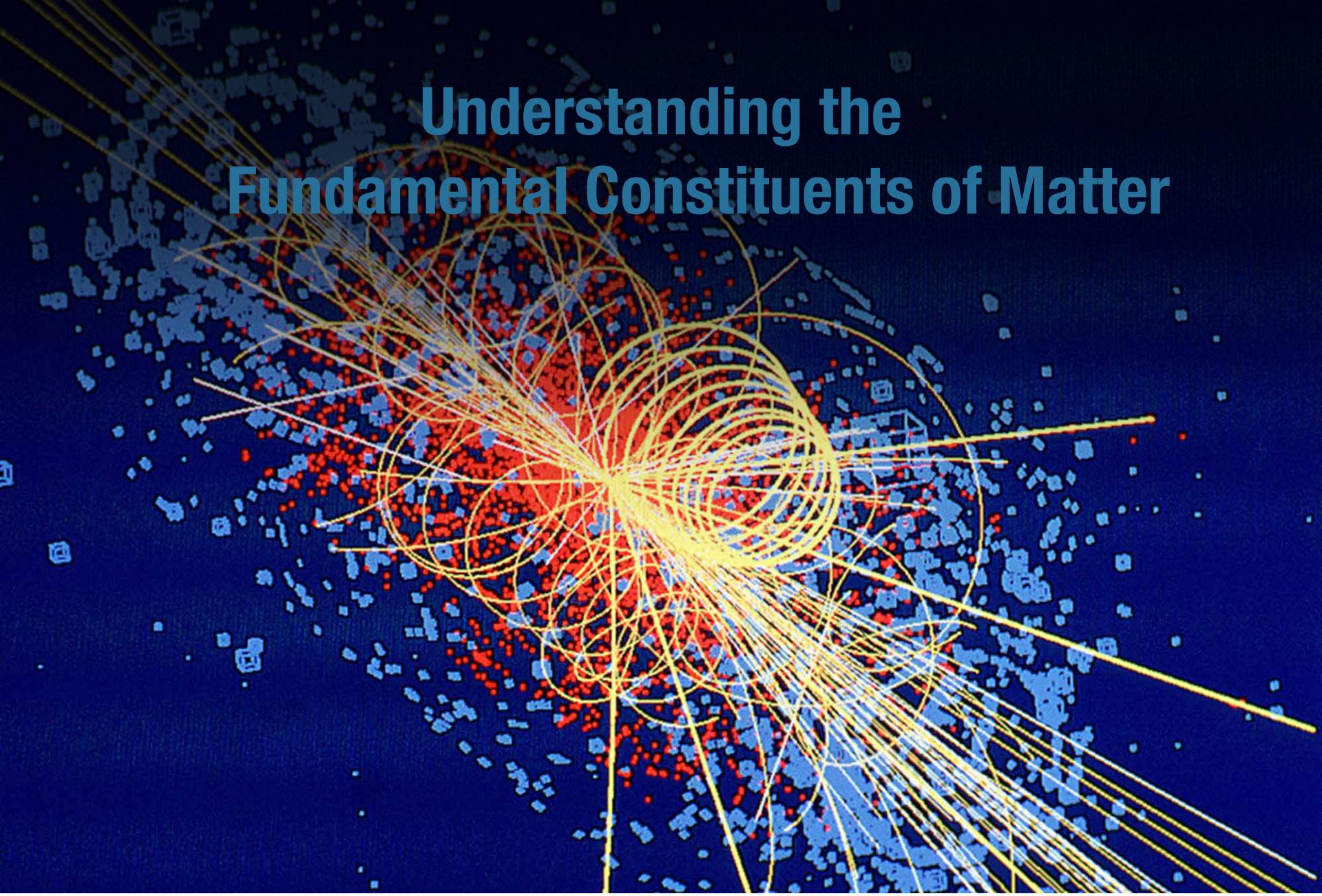
DisCo

- Adam Rupe, Jim Crutchfield (UC Davis); Nalini Kumar (Intel); Karthik Kashinath (NERSC); et al.
- Unsupervised pattern *discovery* technique based on computational mechanics; capable of recovering coherent flow structures
- Implemented in Python





Understanding the Fundamental Constituents of Matter

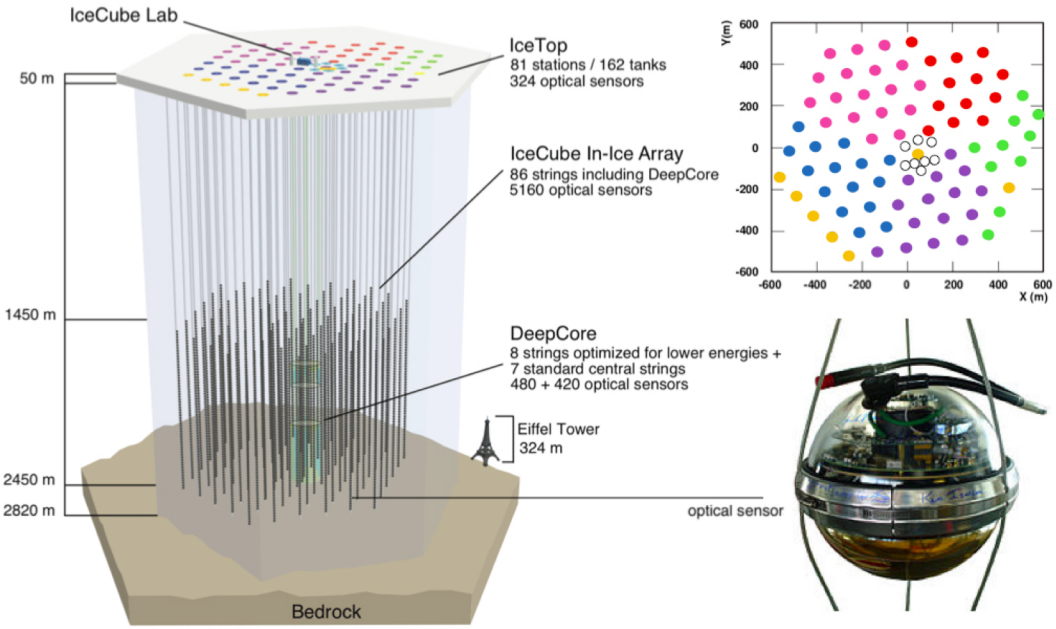


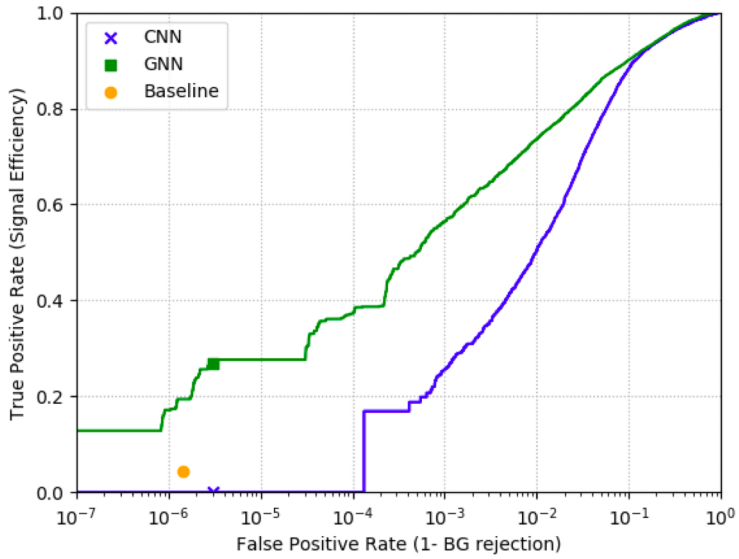
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Graph NNs for Neutrino classification

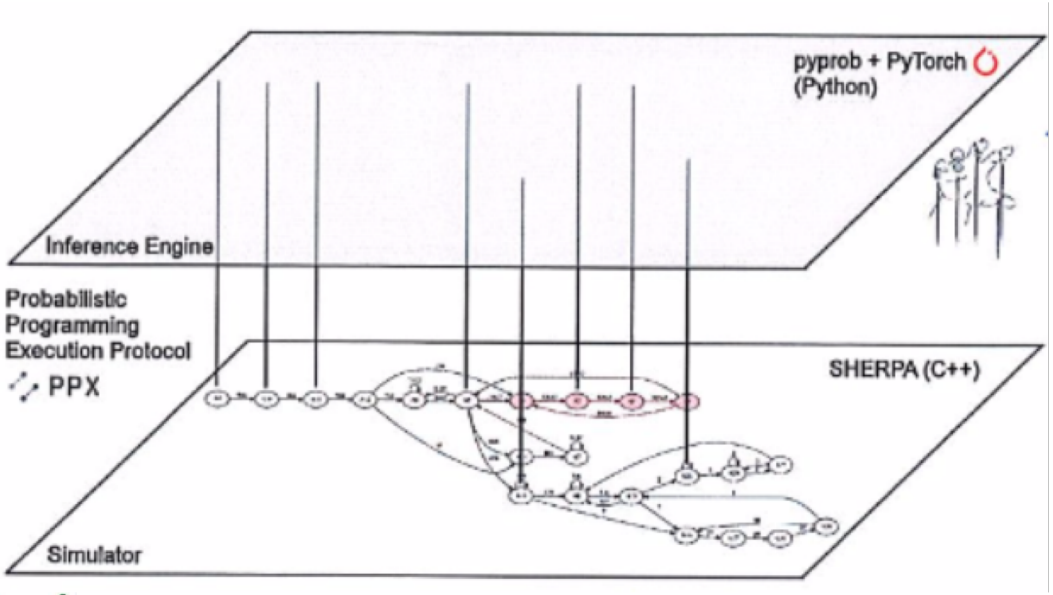
- Nick Choma, Joan Bruna (NYU); Federico Monti, Michael Bronstein (ICS, U. Svizzera); Spencer Klei, Tomasz Palczewski (LBL Physics); Lisa Gerhardt, Wahid Bhimji (NERSC)



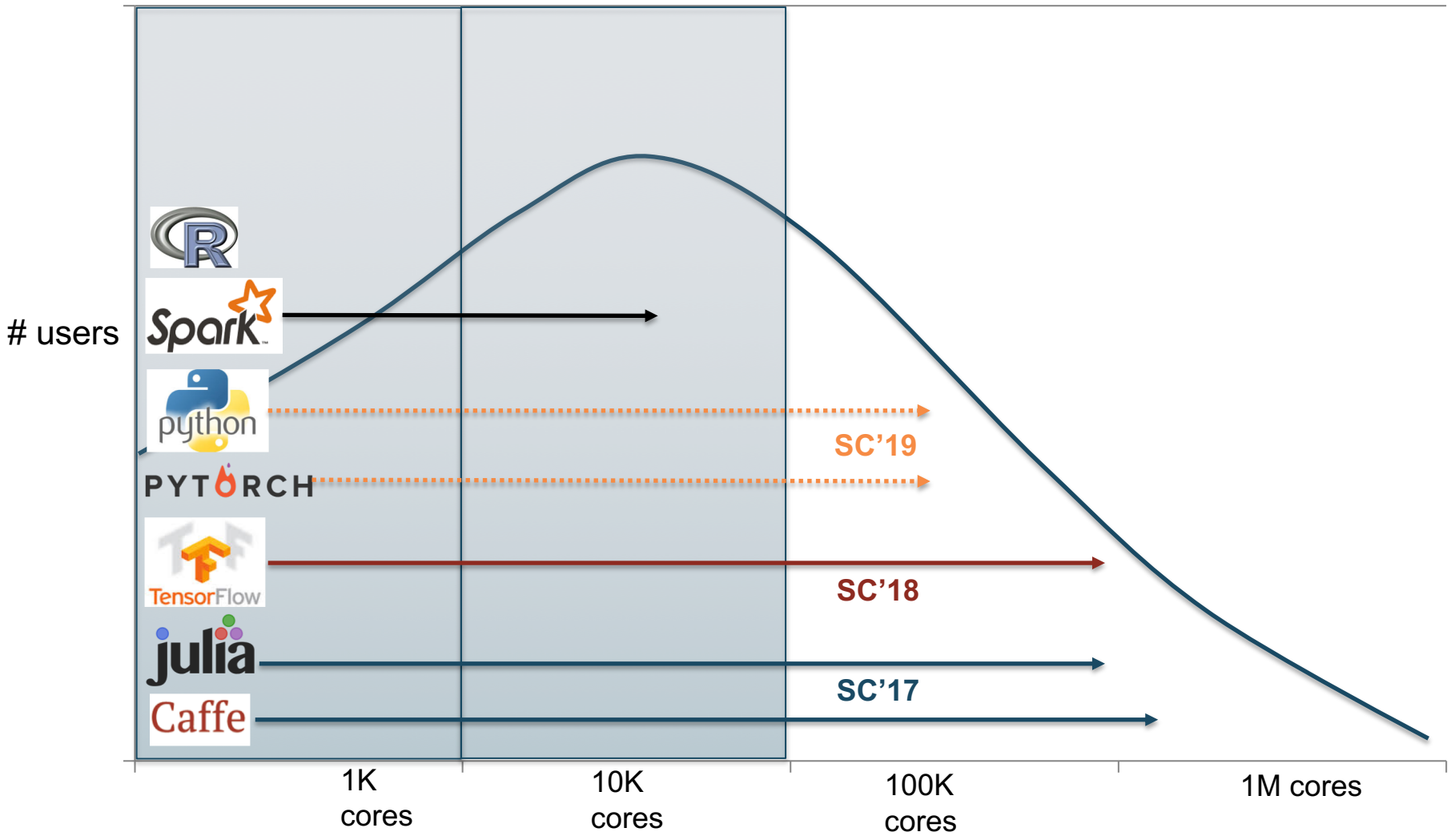


etalumis

- Frank Wood (UBC); Gunes Baydin (Oxford); Kyle Cranmer (NYU); Lei Shao (Intel); Wahid Bhimji (NERSC); et al.
- Goal: Pattern classification and anomaly detection on HEP data
- Approach: Combine Probabilistic Programming and Deep Learning
- Dynamic, hybrid 3D Conv + LSTM architecture



Data Analytics on Cori



Broader Deployment / Availability

- Performance enhancements incorporated into builds of Julia, Caffe, TensorFlow, MKL-DNN, MLSL, Cray DL plugin, …

- NERSC software builds and documentation has been updated

- Select applications being converted to benchmarks
  - Deep Learning benchmark suite for analytics
  - Celeste I/O benchmark
  - ‘Workflow’ benchmark

Conclusions

- Big Data Center collaboration is enabling capability applications on Cori
  - NERSC: Domain Science drivers in Astronomy, Cosmology, Climate and HEP
  - IPCCs: State-of-the-art research in Deep Learning, Graphical Models, Probabilistic Programming, Topological Methods and Computational Mechanics
  - Intel and Cray: Performance Optimization and Scaling of Production Data Analytics and Management technologies

- We intend to share lessons learnt, code, benchmarks with the broader community

- We welcome collaborations!

Thank You!


NERSC


U.S. DEPARTMENT OF
ENERGY
Office of
Science


BERKELEY LAB